# Noise Covariance Matrix Estimation for Rotating Microphone Arrays

Alastair H. Moore, *Member, IEEE*, Wei Xue, *Member, IEEE*, Patrick A. Naylor, *Senior Member, IEEE*, and Mike Brookes, *Member, IEEE*

***Abstract*—The noise covariance matrix computed between the signals from a microphone array is used in the design of spatial filters and beamformers with applications in noise suppression and dereverberation. This paper specifically addresses the problem of estimating the covariance matrix associated with a noise field when the array is rotating during desired source activity, as is common in head-mounted arrays. We propose a parametric model that leads to an analytical expression for the microphone signal covariance as a function of the array orientation and array manifold. An algorithm for estimating the model parameters during noise-only segments is proposed and the performance shown to be improved, rather than degraded, by array rotation. The stored model parameters can then be used to update the covariance matrix to account for the effects of any array rotation that occurs when the desired source is active. The proposed method is evaluated in terms of the Frobenius norm of the error in the estimated covariance matrix and of the noise reduction performance of a minimum variance distortionless response beamformer. In simulation experiments the proposed method achieves 18 dB lower error in the estimated noise covariance matrix than a conventional recursive averaging approach and results in noise reduction which is within 0.05 dB of an oracle beamformer using the ground truth noise covariance matrix.**

***Index Terms*—Covariance matrix estimation, spatial filtering, spherical harmonic analysis, adaptive estimation, moving microphone array.**

## I. Introduction

SPATIAL filtering is a fundamental tool for multichannel signal enhancement in noisy and reverberant environments and is used in many applications, such as telecommunications, automatic speech recognition, human-robot interaction, assistive listening devices and hearing aids. The widely used minimum variance distortionless response (MVDR) beamformer [1] requires knowlege of two quantities: the steering vector, which defines the distortionless constraint, and the noise covariance matrix (NCM), which describes the interchannel relationship of the undesired signal. The focus of this contribution is the estimation of the NCM encountered by a microphone array in a non-isotropic sound field when the array can rotate freely in three dimensions.

Estimation of the NCM is required both for MVDR beamforming and also in multichannel Wiener filter (MWF)-based enhancement [2]. A direct implementation of the MWF requires the desired source covariance matrix in addition to the NCM. Alternatively, it has been shown that the MWF is equivalent to an MVDR beamformer followed by a Wiener post-filter [3]. In this case knowlege of the NCM improves the estimate of the signal-to-noise ratio (SNR) [4]–[8] or coherent-to-diffuse ratio [9]–[14], which is used to calculate the post-filter gain. It is common to calculate the NCM based on an assumed model of the noise field. Commonly-used models are spatially white noise [4], [5], spherically isotropic noise [6], [7] or cylindrically isotropic noise [14]–[16]. These models do not account for the true spatial distribution of the acoustic noise field since they are independent of the observed microphone signals.

Adaptive estimation of the NCM normally requires noise-only segments to be identified or an estimate of the speech-absence probability to be determined [17]. In [18], for example, it is assumed that the spatial characteristics of the noise do not change while the desired source is active. This allows an MVDR beamformer designed during noise-only segments to be used during speech activity.

A major source of non-stationarity in the NCM which arises in real-world situations is due to movement of the microphone array. We consider in particular the case of array rotation, which is common in situations where microphones are mounted on the head of a person, for example as part of an assisted listening system, or of a robot, for example for human-robot interaction. In such scenarios it is typical for the array to rotate in response to a desired source. For example, the robot or human listener might turn in order to face a talker that has just started speaking or turn in response to an instruction to look in a particular direction. Head movements are also an integral part of nonverbal communications [19] and so should be expected in real-world listening. Crucially, in these contexts, the desired source is active during the array rotation, preventing an immediate update of the NCM estimate.

Previous work on rotating microphone arrays has not explicitly estimated the NCM. Beamforming for rotating microphone arrays is proposed in [20] for a linear array and in [21] for a 2-channel binaural hearing aid array. In that approach only rotations in the horizontal plane are accommodated and, by using

Manuscript received April 11, 2018; revised September 14, 2018 and November 13, 2018; accepted November 13, 2018. Date of publication November 19, 2018; date of current version December 28, 2018. This work was supported by the Engineering and Physical Sciences Research Council under Grant EP/M026698/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sven Nordholm. *(Corresponding author: Alastair H. Moore.)*

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, U.K. (e-mail: alastair.h.moore@imperial.ac.uk; w.xue@imperial.ac.uk; p.naylor@imperial.ac.uk; mike.brookes@imperial.ac.uk).

Digital Object Identifier 10.1109/TASLP.2018.2882307

a generalized sidelobe canceller structure, no estimate of the NCM is obtained.

A number of authors have found it convenient to use a spherical harmonic (SH) description of the plane-wave density (PWD) of a sound field [22] because it is compact and straightforwardly accommodates rotation. In [23], a sound field is sampled at multiple time instants and positions by rotating and/or translating a spherical microphone array. Under the assumption that the sound field is stationary, successive observations can be combined, leading to a higher order SH description of the sound field than could be obtained using a stationary array. The increased spatial resolution enables improved estimation of the direction of arrival (DOA) of the source. A SH sound-field representation is used in [24]–[26] to synthesize binaural microphone signals. The use of spherical microphone arrays to obtain a SH representation of the sound field and to describe its direction dependence in terms of the steered response power is addressed in [27]–[30]. In [31] the SH domain covariance matrix is used to estimate the diffuseness of the sound field which is modelled as an isotropic, and so rotation-invariant, background with individual coherent components. Like [31], we consider the sound field's SH covariance matrix. However, in this work, non-isotropic directionally-uncorrelated sound fields are considered.

The main contribution of this work is a method for estimating the NCM for a rotating microphone array. To this end, supplementary contributions are (i) a spherical harmonic model for the direction dependence of a directionally-uncorrelated noise field, (ii) a simplified analytical expression for the covariance between spherical harmonic coefficients of a directionally-uncorrelated field under the proposed model, and (iii) a derivation of the relationship between the proposed model and the NCM of an arbitrary known microphone array with arbitrary known rotation.

The remainder of the paper is organized as follows. In Section II the problem is formulated. In Section III the notation and some key properties of SH analysis are briefly reviewed. In Section IV the proposed model of the non-isotropic directionally-uncorrelated field is presented and an analytical expression for the resulting NCM is derived. In Section V an algorithm for estimating the parameters of the proposed model is presented. Simulation experiments which confirm the efficacy of the method under ideal and non-ideal conditions are presented in Section VI. Finally, conclusions are drawn in Section VII.

## II. Problem Formulation

We consider a sound field which is sampled in successive time frames by an array of $Q$ microphones. During speech absence, denoted $\mathcal{H}_0 = 1$, the sound field is considered to be noise-only and is assumed to be in the far-field. The power of the noise incident from each direction, or noise power distribution (NPD), is assumed to vary only slowly with time and so, over the time-scales considered in this paper, is treated as constant. Since, in general, the NPD varies with direction, the acoustic NCM of the microphone signals depends on both the free-field array manifold and the orientation, $\Lambda$, of the array. We consider the case where the rotation rate, array dimensions and frame length are such that Doppler effects are insignificant [32]. In addition, we assume that the acoustic noise sound field is directionally-uncorrelated and neglect correlations arising from multipath. The goal of this work is to estimate the time-varying NCM from knowledge of the microphone signals, the time-varying array orientation and the free-field array manifold.

Rather than estimate the NCM directly it is proposed to estimate the parameters of a model for the NPD since these are independent of array rotation. The estimated NPD is then used to derive the NCM as a function of the known array orientation. An advantage of this approach is that the estimated NCM can be updated in response to array rotation even during speech presence, i.e., $\mathcal{H}_0 = 0$.

The acoustic noise field is described by the PWD, $\underline{a}(\ell, \underline{\Omega})$, where $\underline{\Omega}$ is the direction of incident plane waves in world coordinates. Throughout this paper, we use an underbar to denote quantities that are defined in world coordinates and are therefore unaffected by array rotation. We assume that $\underline{a}(\ell, \underline{\Omega})$ is zero-mean and, over the time scales considered in this paper, wide-sense stationary and ergodic. The NPD, $\underline{s}(\underline{\Omega})$, gives the direction dependence of the noise-field power and is

$$\underline{s}(\underline{\Omega}) = \mathbb{E}\{|\underline{a}(\ell, \underline{\Omega})|^2\} \tag{1}$$

where $\ell$ is the time-frame index and $\mathbb{E}\{\cdot\}$ denotes expectation over realizations of the noise field. Note that, because of the stationarity assumption, $\underline{s}(\underline{\Omega})$ is independent of $\ell$.

The vector of noise signals, $\mathbf{v}(\nu, \ell)$, recorded by an array of $Q$ microphones at frequency index $\nu$ and time-frame index $\ell$ is expressed directly in the short time Fourier transform (STFT) domain

$$\mathbf{v}(\nu, \ell) = \mathbf{x}(\nu, \ell) + \boldsymbol{u}(\nu, \ell) \tag{2}$$

where $\mathbf{x}(\nu, \ell)$ and $\boldsymbol{u}(\nu, \ell)$ are the vectors of $Q$ complex-valued microphone signals due to the acoustic noise and sensor noise, respectively. Since each frequency bin is processed independently, the dependence on $\nu$ is omitted below.

Assuming $\mathbf{x}(\ell)$ and $\boldsymbol{u}(\ell)$ are zero-mean and uncorrelated, the NCM of the microphone noise signals is

$$\mathbf{R}_v(\ell) = \mathbf{R}_x(\ell) + \mathbf{R}_u \tag{3}$$

where $\mathbf{R}_v(\ell) \triangleq \mathbb{E}\left\{\mathbf{v}(\ell)\mathbf{v}^H(\ell)\right\}$ and $(\cdot)^H$ is the conjugate transpose. The acoustic NCM, $\mathbf{R}_x(\ell)$, and sensor NCM, $\mathbf{R}_u$ are similarly defined. Note that, like the NPD, the sensor NCM is assumed to be stationary over the time scales considered. Therefore, the time-dependence in (3) arises only from array rotation.

In Section IV an expression which relates the acoustic NCM to a parametric model of the NPD is developed for a single realization of the array rotation, $\Lambda$. The derivation assumes that noise signals incident from each direction are uncorrelated with each other, i.e., $\underline{s}(\underline{\Omega})$ is directionally-uncorrelated. This assumption will not necessarily be valid if the environment has significant reverberation and the noise arises from a small number of dominant sources. If however the noise arises from a large number of spatially disbursed sources then a directionally-uncorrelated noise field is believed to be a good approximation even in a reverberant environment such as, for example, a crowded restaurant.

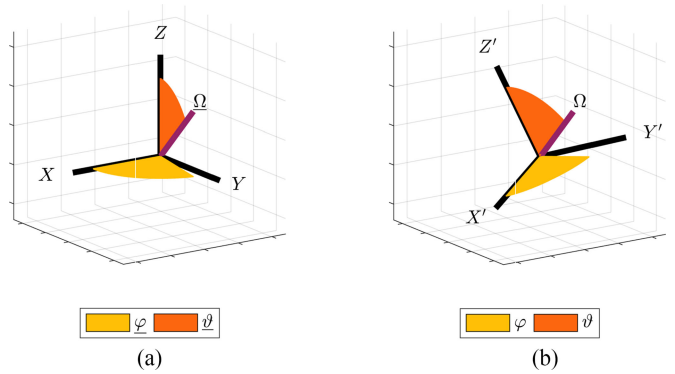



Fig. 1. Azimuth and inclination of a direction vector in terms of (a) reference coordinates and (b) with respect to rotated array.

In Section V the expression developed in Section IV is used with a time-varying array rotation, $\Lambda(\ell)$, to obtain an algorithm to estimate the time-varying NCM, $\mathbf{R}_v(\ell)$.

In this work directions are expressed both in world coordinates and array coordinates. Fig. 1(a) illustrates how an arbitrary direction may be expressed in terms of its inclination, $\underline{\vartheta}$, and azimuth, $\underline{\varphi}$, in world coordinates where $\underline{\Omega} = (\underline{\vartheta}, \underline{\varphi})$. Fig. 1(b) shows the same direction but now expressed in array coordinates $\Omega = (\vartheta, \varphi)$. For an arbitrary array rotation, $\Lambda$, the relation between $\Omega$ and $\underline{\Omega}$ is given in [33, eq 1.65] and is here denoted by $\underline{\Omega}(\Omega, \Lambda)$.

## III. Key Properties of Spherical Harmonics

This section briefly presents the key properties of SH analysis to introduce the required notation. For a comprehensive introduction the reader is referred to [33] or [34].

The complex SH functions of order $n \geq 0$ and degree $m$ with $|m| \leq n$ are defined for $\Omega = (\vartheta, \varphi) \in S^2$, as

$$Y_n^m(\Omega) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\vartheta)\, e^{\imath m\varphi} \tag{4}$$

where $P_n^m(\cdot)$ is the associated Legendre function and $\imath = \sqrt{-1}$. The SH functions form an orthonormal basis where increasing $n$ and $m$ results in functions, $Y_n^m(\Omega)$, with higher spatial frequency. The spherical Fourier transform (SFT) of a square-integrable function, $\mathcal{K}(\Omega)$, is given by

$$\mathcal{K}_{n,m} = \int_{\Omega \in S^2} \mathcal{K}(\Omega)\, [Y_n^m(\Omega)]^*\, d\Omega. \tag{5}$$

Assuming $\mathcal{K}(\Omega)$ is spatially bandlimited, its order, $N_\mathcal{K}$, is the maximum $n$ for which any $\mathcal{K}_{n,m} > 0$.

To avoid a plethora of subscripts, we index the individual SH functions with the single index $p = n^2 + n + m + 1$ such that $Y_p(\Omega) \equiv Y_n^m$ and $\mathcal{K}_p \equiv \mathcal{K}_{n,m}$, where $1 \leq p \leq P_\mathcal{K} = (N_\mathcal{K} + 1)^2$. The inverse index mappings are given by $n(p) = \lfloor \sqrt{p-1} \rfloor$ and $m(p) = p - n(p)^2 - n(p) - 1$ where $\lfloor \cdot \rfloor$ denotes the floor function.

The inverse SFT (ISFT) decomposes $\mathcal{K}(\Omega)$ in terms of the $Y_p(\Omega)$ and may be written in vector form as

$$\mathcal{K}(\Omega) = \boldsymbol{\kappa}^T \mathbf{y}(\Omega) \tag{6}$$

where the SH coefficient vector $\boldsymbol{\kappa} = [\, \mathcal{K}_1 \ \ldots \ \mathcal{K}_{P_\mathcal{K}} \,]^T$, the SH function vector $\mathbf{y}(\Omega) = [\, Y_1(\Omega) \ \ldots \ Y_{P_\mathcal{K}}(\Omega) \,]^T$ and $(\cdot)^T$ denotes the transpose.

In practical applications $\mathcal{K}_p$ is obtained by sampling $\mathcal{K}(\Omega)$ at $I$ discrete angles, $\Omega_i$,

$$\mathcal{K}_p = \sum_{i=1}^{I} \varpi_{p,i} \mathcal{K}(\Omega_i) Y_p^*(\Omega_i) \tag{7}$$

where the quadrature weights, $\varpi_{p,i}$, are chosen to ensure the orthonormality condition of the spherical harmonics, i.e.,

$$\sum_{i=1}^{I} \varpi_{p,i} Y_p(\Omega_i) Y_{p'}^*(\Omega_i) = \delta_{p,p'} \tag{8}$$

where $\delta_{a,b} = 1$ when $a = b$ and 0 otherwise. The maximum spatial frequency which can be sampled without spatial aliasing is determined by the angular arrangement of sample points. For a function of order $N_\mathcal{K}$ the sampling scheme requires at least $I \geq (N_\mathcal{K} + 1)^2$ directions to be sampled. If $I$ is strictly greater than this bound, the grid is oversampled and the weights are not unique; the choice of weights in this case is discussed in Chapter 3 of [33].

A delta function on the sphere at $\Omega' = (\vartheta', \varphi')$ is denoted

$$\mathcal{K}^{\dagger(\Omega')}(\Omega) = \delta(\Omega - \Omega') \tag{9}$$

$$= \delta(\cos\vartheta - \cos\vartheta')\delta(\varphi - \varphi') \tag{10}$$

and its SFT is $\mathcal{K}_p^{\dagger(\Omega')} = Y_p(\Omega'), \forall p$. Since the SH coefficients do not decay in amplitude with $p$, it follows that $\mathcal{K}^{\dagger(\Omega')}(\Omega)$ has infinite spatial bandwidth.

If $\mathcal{K}(\Omega)$ is expressed in a rotated frame of reference as $\underline{\mathcal{K}}(\underline{\Omega}(\Omega, \Lambda))$, then the resultant SH coefficients may be obtained in the SH domain as

$$\underline{\boldsymbol{\kappa}} = \mathbf{D}(\Lambda)\boldsymbol{\kappa} \tag{11}$$

where $\mathbf{D}(\Lambda)$ is the Wigner-D rotation matrix [33], which is block-diagonal and sparse. The notational simplicity of rotation in the SH domain along with the computational benefits of the sparsity of $\mathbf{D}(\Lambda)$ make it appealing for describing the signals encountered by a microphone array under rotation.

## IV. Proposed Parametric Model

In Section IV-A the acoustic NCM is related to the covariance matrix of a SH domain representation of the PWD of the sound field. In Section IV-B a parametric model for the NPD is presented and its relationship to the acoustic NCM derived. Throughout this section, for clarity of notation, the microphone array orientation, $\Lambda$, is taken to have a constant value which, in turn, means that the acoustic NCM, $\mathbf{R}_x$, is time-invariant.

### A. Acoustic NCM From Sound-Field Covariance

Assuming the effective support of the array manifold is short compared to the STFT frame length, the acoustic noise, $x_q(\ell)$, observed by the $q^{\text{th}}$ microphone may be expressed as the array

response to an infinite sum of plane waves over $S^2$

$$x_q(\ell) = \int_{\Omega\in S^2} h_q(\Omega)\underline{a}\big(\ell, \underline{\Omega}(\Omega,\Lambda)\big)d\Omega \tag{12}$$

where $\underline{a}(\ell,\underline{\Omega})$ is expressed in world coordinates and represents the PWD of the sound field at the origin in the absence of the microphone array and $h_q(\Omega)$ for $1 \le q \le Q$ is the array manifold.

In order to express (12) in terms of SHs, let $\mathbf{a}(\ell)$ be the SH coefficient vector representing the PWD of the sound field in array coordinates. In world coordinates, $\underline{a}(\ell,\underline{\Omega}(\Omega,\Lambda)) = a\big(\ell,\Omega\big)$ may be decomposed using the ISFT from (6)

$$\underline{a}\big(\ell,\underline{\Omega}(\Omega,\Lambda)\big) = a(\ell,\Omega) \tag{13}$$

$$= \mathbf{a}^T(\ell)\mathbf{y}(\Omega). \tag{14}$$

To express $h_q(\Omega)$ in the SH domain ammenable to simplification, the SFT of the conjugate array manifold, $h_q^*(\Omega)$, of microphone $q$ is defined as [24]

$$\tilde{h}_{q,p} = \int_{\Omega\in S^2} h_q^*(\Omega)Y_p^*(\Omega)d\Omega \tag{15}$$

with the corresponding ISFT as

$$h_q^*(\Omega) = \tilde{\mathbf{h}}_q^T\mathbf{y}(\Omega) \tag{16}$$

where $\tilde{\mathbf{h}}_q = \big[\,\tilde{h}_{q,1}\ \ldots\ \tilde{h}_{q,P_h}\,\big]^T$. Note that properties of the SFT imply that, in general, $\tilde{h}_{q,p}^* \neq h_{q,p}$.

Substituting (14) and the conjugate of (16) into (12) gives

$$x_q(\ell) = \int_{\Omega\in S^2} \tilde{\mathbf{h}}_q^H\mathbf{y}^*(\Omega)\mathbf{y}^T(\Omega)\mathbf{a}(\ell)d\Omega \tag{17}$$

$$= \tilde{\mathbf{h}}_q^H\left(\int_{\Omega\in S^2}\mathbf{y}^*(\Omega)\mathbf{y}^T(\Omega)d\Omega\right)\mathbf{a}(\ell) \tag{18}$$

$$= \tilde{\mathbf{h}}_q^H\mathbf{a}(\ell) \tag{19}$$

where the simplification in (19) follows from the orthonormality of SHs [33]. The vector, $\mathbf{x}(\ell)$, of all $Q$ microphone signals is therefore given by

$$\mathbf{x}(\ell) = \widetilde{\mathbf{H}}^H\mathbf{a}(\ell) \tag{20}$$

where $\widetilde{\mathbf{H}} = \big[\,\tilde{\mathbf{h}}_1\ \tilde{\mathbf{h}}_2\ \ldots\ \tilde{\mathbf{h}}_Q\,\big]$.

The acoustic NCM can therefore be expressed as

$$\mathbf{R}_x = \widetilde{\mathbf{H}}^H\mathbf{R}_a\widetilde{\mathbf{H}} \tag{21}$$

where

$$\mathbf{R}_a = \mathbb{E}\left\{\mathbf{a}(\ell)\mathbf{a}^H(\ell)\right\} \tag{22}$$

is the SH covariance matrix of the acoustic noise field in array coordinates. Expressed in vectorized form [35], (21) is

$$\mathbf{r}_x = \left(\widetilde{\mathbf{H}}^T\otimes\widetilde{\mathbf{H}}^H\right)\mathbf{r}_a \tag{23}$$

where $\mathbf{r}_x = \overrightarrow{\mathbf{R}_x}$, $\mathbf{r}_a = \overrightarrow{\mathbf{R}_a}$, $\otimes$ denotes the Kronecker product and $\overrightarrow{\cdot}$ denotes the vectorization of a matrix obtained by concatenating its columns. Thus (21) and (23) express the acoustic NCM in terms of the SH covariance matrix of the acoustic noise field.

### B. Spherical Harmonic Model of Noise Power Distribution

If the acoustic noise sound field is zero-mean and directionally-uncorrelated, the covariance between two directions, $\underline{\Omega}$ and $\underline{\Omega}'$, of the sound-field PWD is

$$\mathbb{E}\left\{\underline{a}(\ell,\underline{\Omega})\underline{a}^*(\ell,\underline{\Omega}')\right\} = \underline{s}(\underline{\Omega})\delta(\underline{\Omega}-\underline{\Omega}') \tag{24}$$

where $\underline{s}(\underline{\Omega})$ is the noise power distribution (NPD) defined in (1). The parameters of the proposed model are the SFT coefficients, $\underline{\mathbf{s}}$, of $\underline{s}(\underline{\Omega})$ which satisfy the ISFT relation

$$\underline{s}(\underline{\Omega}) = \underline{\mathbf{s}}^T\mathbf{y}(\underline{\Omega}) = \mathbf{y}^T(\underline{\Omega})\underline{\mathbf{s}} \tag{25}$$

where $\underline{\mathbf{s}}$ is a $P_s$ element vector of SH coefficients. The valid range of $\underline{\mathbf{s}}$ is constrained such that the NPD, $\underline{s}(\underline{\Omega})$, is real-valued and non-negative for all $\underline{\Omega}$.

The model parameters, $\underline{\mathbf{s}}$, are defined in world coordinates which means that they are independent of the array rotation, $\Lambda$. In developing an expression for the acoustic NCM, $\mathbf{R}_x$, it is useful to express the NPD in array coordinates, $s(\Omega)$. Using (11) and (25), $s(\Omega)$ can be written as

$$s(\Omega) = \underline{s}(\underline{\Omega}(\Omega,\Lambda)) \tag{26}$$

$$= \mathbf{y}^T(\Omega)\mathbf{D}(\Lambda^{-1})\underline{\mathbf{s}} \tag{27}$$

where $\Lambda^{-1}$ denotes the inverse rotation of $\Lambda$.

Using (27), it is shown in Appendix VII that element $(p',p'')$ of $\mathbf{R}_a$ in (22) is given by

$$\mathbb{E}\left\{a_{p'}(\ell)a_{p''}^*(\ell)\right\} = \mathbf{g}_{p',p''}^T\mathbf{D}^T(\Lambda)\underline{\mathbf{s}}^* \tag{28}$$

where $\mathbf{g}_{p',p''} = \big[\,G_{1,p',p''}\ \ldots\ G_{P_s,p',p''}\,\big]^T$ and

$$G_{p,p',p''} = \int_{\Omega\in S^2} Y_p(\Omega)Y_{p'}(\Omega)Y_{p''}^*(\Omega)d\Omega \tag{29}$$

is the Gaunt coefficient, for which a closed form solution is given in [36, pp. 39–40]. Therefore the vectorized SH covariance matrix, $\mathbf{r}_a$ from (23), of a directionally-uncorrelated sound field can be written as

$$\mathbf{r}_a = \mathbf{G}\mathbf{D}^T(\Lambda)\underline{\mathbf{s}}^* \tag{30}$$

where $\mathbf{G}$ is a $P_h^2 \times P_s$ matrix in which row $p' + (p''-1)P_h$ is equal to $\mathbf{g}_{p',p''}^T$ and $P_h$ and $P_s$ are the number of SH coefficients used to describe the microphone array manifold, $h_q(\Omega)$, and the noise power distribution (NPD), $\underline{s}(\underline{\Omega})$, respectively. The choice of $P_h$ can be determined a priori since it depends only on the microphone array geometry. The effect of this choice for an illustrative array is shown in Experiment 3. The choice of $P_s$ depends on the desired spatial resolution of the NPD; the effect of mismatch is examined in Experiment 4.

Using (30), the SH covariance matrix of a directionally-uncorrelated sound field can be described in terms of a weighted sum of analytically defined matrices. Fig. 2 illustrates the columns of $\mathbf{G}$ corresponding to SH bases up to $n = 2$, arranged as $P_h \times P_h$ matrices. For the special case of a spherically isotropic sound field, only the first element of $\underline{s}$ is non-zero. As illustrated in the top row of Fig. 2, in this case, $\mathbf{R}_a$ reduces to a scaled identity matrix, as has been shown in [31], [37], [38]. In general it can be seen that $\mathbf{G}$ is very sparse and that the magnitudes of its elements do not decay as $p'$ and $p''$ increase.

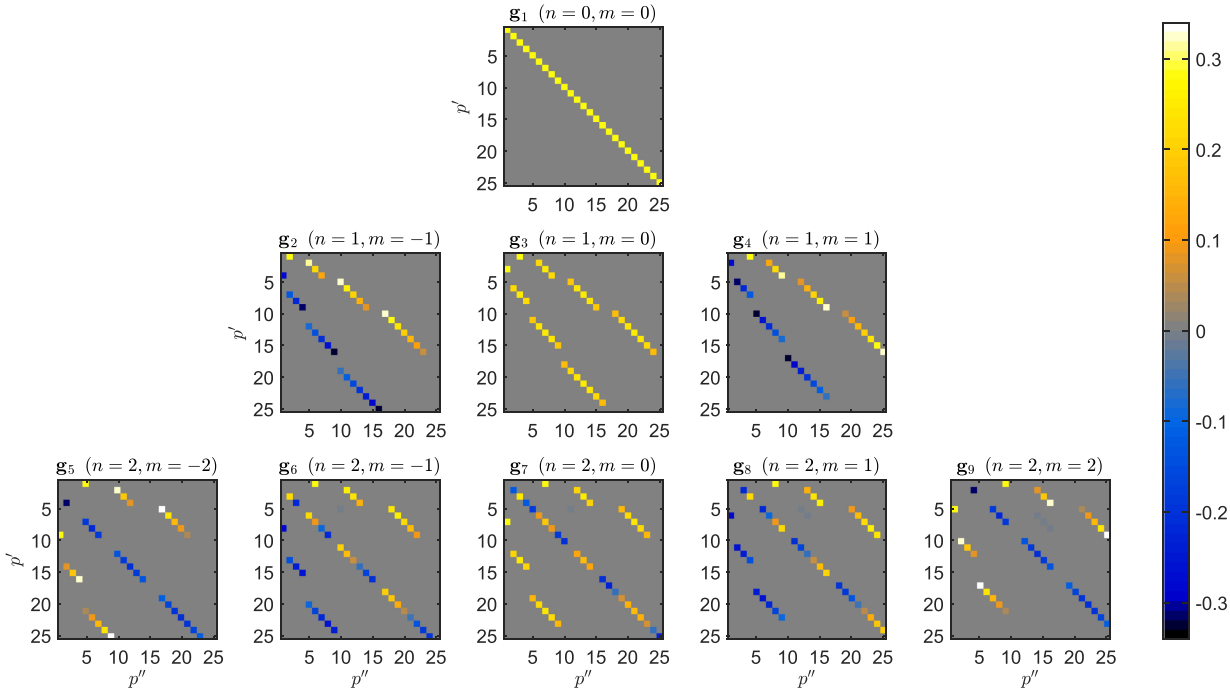



Fig. 2. Examples of Gaunt coefficients arranged as matrices according to proposed SH parameterization of directionally-uncorrelated sound-field power distribution.

Substituting (30) into (23) the contribution to the acoustic NCM is

$$\mathbf{r}_x = \left(\widetilde{\mathbf{H}}^T \otimes \widetilde{\mathbf{H}}^H\right) \mathbf{G}\mathbf{D}^T(\Lambda)\underline{\mathbf{s}}^* \tag{31}$$

$$= \mathbf{B}\mathbf{D}^T(\Lambda)\underline{\mathbf{s}}^* \tag{32}$$

where

$$\mathbf{B} = \left(\widetilde{\mathbf{H}}^T \otimes \widetilde{\mathbf{H}}^H\right) \mathbf{G}. \tag{33}$$

Note that $\mathbf{B}$ is a $Q^2 \times P_s$ matrix and is independent of both array rotation and fluctuations in the sound field. This time-invariance means it can be calculated once for a given array manifold. As a result, for a given array manifold, the cost of calculating $\mathbf{R}_x$ from (32) is independent of $P_h$, the number of SH coefficients used to describe the array manifold.

## V. Model Parameter Estimation

In (32) a mapping between the proposed model parameters, $\underline{\mathbf{s}}$, and the acoustic NCM, $\mathbf{R}_x$, was established. We now propose an online algorithm based on exponentially-weighted least squares (EWLS) for estimating $\mathbf{R}_x$ and $\mathbf{R}_u$ jointly. Together, these quantities determine the total NCM, $\mathbf{R}_v$, which is ultimately required in order, for example, to solve for the MVDR beamformer weights. It is assumed that the NPD SH coefficients, $\underline{\mathbf{s}}$, to be estimated are static or only slowly varying. Estimated quantities are denoted $\hat{\cdot}$ and are time varying with each signal frame, $\ell$, such that

$$\hat{\mathbf{r}}_v(\ell) = \hat{\mathbf{r}}_x(\ell) + \hat{\mathbf{r}}_u(\ell). \tag{34}$$

From (32) the estimate of $\mathbf{r}_x(\ell)$ is given by

$$\hat{\mathbf{r}}_x(\ell) = \mathbf{B}\mathbf{D}^T(\Lambda(\ell))\hat{\underline{\mathbf{s}}}^*(\ell) \tag{35}$$

where the dependence of $\Lambda(\ell)$ on $\ell$ is explicit. The error introduced by frame-based processing of signals observed by an array during rotation is dependent on the radius of the array, frame length, signal frequency and rate of rotation and is discussed in detail in [32]. In this paper, we assume that the combination of these factors is chosen such that this source of error can be neglected.

The variance of the spatially white noise at the $q^{\text{th}}$ microphone is $\phi_q$ leading to $\mathbf{R}_u = \text{diag}(\boldsymbol{\phi})$ where $\boldsymbol{\phi} = [\,\phi_1 \ \ldots \ \phi_Q\,]^T$ and the diagonal operator, $\text{diag}(\cdot)$, gives a diagonal matrix. In vectorized form, the estimated sensor NCM, $\hat{\mathbf{r}}_u(\ell)$, is related to the estimated parameters, $\hat{\boldsymbol{\phi}}(\ell)$, as

$$\hat{\mathbf{r}}_u(\ell) = \mathbf{M}\hat{\boldsymbol{\phi}}(\ell) \tag{36}$$

where $\mathbf{M} = [\overrightarrow{\text{diag}(\wedge_1)}\ \overrightarrow{\text{diag}(\wedge_2)}\ \ldots\ \overrightarrow{\text{diag}(\wedge_Q)}]$ and $\wedge_q$ has a one in the $q^{\text{th}}$ element and zeros elsewhere. In the special case where the spatially white noise power is the same at all microphones, i.e., $\phi_q = \phi,\ \forall q$, (36) simplifies to

$$\hat{\mathbf{r}}_u(\ell) = \overrightarrow{\mathbf{I}}\,\hat{\phi}(\ell) \tag{37}$$

where $\mathbf{I}$ is the $Q \times Q$ identity matrix.

Combining (35) and (36) the estimate of the vectorized NCM is

$$\hat{\mathbf{r}}_v(\ell) = \mathbf{C}(\ell)\hat{\boldsymbol{\theta}}^*(\ell) \tag{38}$$

where

$$\mathbf{C}(\ell) = \left[\mathbf{B}\mathbf{D}^T(\Lambda(\ell))\ \mathbf{M}\right] \tag{39}$$

and $\hat{\boldsymbol{\theta}}(\ell) = \left[\underline{\hat{\mathbf{s}}}^T(\ell)\ \hat{\boldsymbol{\phi}}^T(\ell)\right]^T$.

In order to apply EWLS, we define the instantaneous error at frame $\ell$ as a vector

$$\mathbf{e}(\ell) = \mathbf{d}(\ell) - \hat{\mathbf{r}}_v(\ell) \tag{40}$$

where $\mathbf{d}(\ell) = \overrightarrow{\mathbf{v}(\ell)\mathbf{v}^H(\ell)}$.

The EWLS cost function to be minimized is [39]

$$\xi(\ell) = \sum_{\ell'=1}^{\ell} \lambda^{\ell-\ell'}\mathbf{e}^H(\ell)\mathbf{e}(\ell) \tag{41}$$

where $0 < \lambda^{\ell-\ell'} \leq 1$ is an exponential weighting factor. The optimal parameters, $\hat{\boldsymbol{\theta}}(\ell)$, are the solution to the normal equations

$$\boldsymbol{\Gamma}(\ell)\hat{\boldsymbol{\theta}}(\ell) = \boldsymbol{\rho}(\ell) \tag{42}$$

$$\Rightarrow \quad \hat{\boldsymbol{\theta}}(\ell) = \boldsymbol{\Gamma}^{-1}(\ell)\boldsymbol{\rho}(\ell) \tag{43}$$

where $\boldsymbol{\Gamma}(\ell)$ and $\boldsymbol{\rho}(\ell)$ are the exponentially-weighted autocorrelation matrix and crosscorrelation vector, respectively. During noise-only frames, denoted $\mathcal{H}_0(\ell) = 1$, these are estimated as

$$\boldsymbol{\Gamma}(\ell) = \sum_{\ell'=1}^{\ell} \lambda^{\ell-\ell'}\mathbf{C}^T(\ell')\mathbf{C}^*(\ell') \tag{44}$$

$$= \lambda\boldsymbol{\Gamma}(\ell-1) + \mathbf{C}^T(\ell)\mathbf{C}^*(\ell) \tag{45}$$

and

$$\boldsymbol{\rho}(\ell) = \lambda\boldsymbol{\rho}(\ell-1) + \mathbf{C}^T(\ell)\mathbf{d}^*(\ell). \tag{46}$$

During source activity, $\mathcal{H}_0(\ell) = 0$ and estimates are not updated, i.e.,

$$\boldsymbol{\Gamma}(\ell) = \boldsymbol{\Gamma}(\ell-1) \tag{47}$$

$$\boldsymbol{\rho}(\ell) = \boldsymbol{\rho}(\ell-1). \tag{48}$$

The complete NCM estimation algorithm is summarized in Figure 3.

**for** each frequency bin, $\nu$, **do**
  Determine $\mathbf{B}$ using (33)
  Initialize $\mathbf{C}(0)$ using (39), $\boldsymbol{\Gamma}(0) \leftarrow \mathbf{0}$ and $\boldsymbol{\rho}(0) \leftarrow \mathbf{0}$
  **for** each $\ell$ **do**
    **if** $\Lambda(\ell) \neq \Lambda(\ell-1)$ **then**
      Update $\mathbf{C}$ using (39)
    **if** $\mathcal{H}_0 = 1$ (speech absence) **then**
      Update $\boldsymbol{\Gamma}(\ell)$ using (45)
      Update $\boldsymbol{\rho}(\ell)$ using (46)
      Update $\hat{\boldsymbol{\theta}}(\ell)$ using (43)
    Update $\hat{\mathbf{r}}_v(\ell)$ using (38)

Fig. 3. Algorithm for estimating NCM.

## VI. Evaluation

In this section, simulation experiments are reported which demonstrate the efficacy of the proposed method. Experiment 1 investigates the effect of different rotation sequences on the accuracy of estimating the model parameters and the NCM. Experiment 2 demonstrates the convergence of the estimated NCM compared to conventional signal dependent and independent methods, highlighting in particular the case when array rotation is in response to desired source activity. Experiment 3 analyses the sensitivity of the proposed method to smoothing of the array manifold due to limited density of spatial sampling. Finally, Experiment 4 investigates the impact of varying the number of microphones and of mismatch in the order of the model, $N_{\hat{s}}$, compared to the order of the NPD, $N_s$.

The complete source code required to reproduce the reported experiments is publicly available with DOI 10.5281/zenodo.1410457.

### A. Experiment Setup

The array manifold in (12), $h_q(\nu, \Omega)$, for an array of microphones on the surface of a rigid sphere with radius 9 cm is calculated analytically using a SH expansion [22], [40]. The expansion order is set to 16, which ensures the worst case reconstruction error across all frequencies considered is less than $-80$ dB. The microphones are equally spaced on a circle, $20^\circ$ above the horizontal plane. Experiments 1 to 3 use $Q = 4$ microphones while Experiment 4 uses both $Q = 4$ and also $Q = 16$ to investigate the effect of varying $Q$.

A sound field with known spatial distribution is simulated directly in the time-frequency (TF) domain using independent zero-mean circularly-symmetric Guassian noise signals incident from 578 directions. These directions form a spherical sampling quadrature grid supporting SH decomposition up to order 16. Using (12) discretized according to the quadrature grid, the power of each plane wave is given by $s(\Omega_i)$, as in (27).

Sensor noise is simulated by adding independent zero-mean circularly-symmetric Gaussian noise to each microphone signal. The sensor noise power at each microphone is drawn from a Gaussian distribution with mean fixed at $-20$ dB with respect to the acoustic noise power, averaged over all microphones, and variance equal to 10% of the mean. The simulations therefore represent a typical use case where the received signals are dominated by acoustic noise and the sensor noise is similar, but not identical, across microphones.

Rotation of the array is implemented according to piecewise-constant trajectories so that frames in which a change in the array orientation occurs are clearly identifiable. Details of these trajectories are given in the relevant experiment descriptions. The proposed method requires a measurement of the array orientation at each frame. Errors in the yaw, pitch and roll components of the measured array orientation are simulated as independent identically-distributed zero-mean, additive Gaussian noise with standard deviation, $\sigma_{\mathrm{IMU}} = 1^\circ$.

The power distribution of the synthesized sound fields have an axially-symmetric cardioid shape

$$\underline{s}(\underline{\Omega}) = (1/2)^{N_S}\,(1+\cos(\underline{\Omega}-\underline{\Omega}_0))^{N_S} \tag{49}$$

where $\underline{\Omega}_0$ is the direction of the maximal response and $N_s$ is the order, where higher-order cardioids concentrate the energy over a narrower region. In Experiments 1–3, $N_s = 2$ and, in Experiment 4, $N_s$ is varied from 1 to 4. In each experiment, evaluations are conducted for 20 different sound fields, each with $\underline{\Omega}_0$ aligned to one of the faces of an icosahedron. An oracle voice activity detector (VAD) is used to determine the speech absence state, $\mathcal{H}_0(\ell)$.

The implementation of the proposed method assumes the sensor noise power is the same for all microphones, as in (37), which intentionally introduces a mismatch compared with the simulated conditions. In Experiments 1–3, the SH order of the cardioid sound field and the order of the estimated model are matched $N_{\hat{s}} = N_s = 2$ (from which $P_{\hat{s}} = (N_{\hat{s}}+1)^2 = 9$). Except where otherwise stated, $N_h = 15$ (from which $P_h = (N_h+1)^2 = 256$), the frequency is 2200 Hz and $(1-\lambda) = 1\times 10^{-5}$. Further implementation details, specfic to each experiment, are described below.

### B. Metrics and Baseline Approaches

The error in each of the estimated model parameters is $\epsilon_p(\ell) = |\underline{s}_p - \hat{\underline{s}}_p(\ell)|/|\underline{s}_1|$, where $|\underline{s}_1|$ is used as the normalizing factor in the denominator because it is constant for all sound fields tested whereas $|\underline{s}_p|$ is zero in some cases.

The error in the estimated NCM is assessed as the Frobenius norm of the scale-invariant error

$$\mathcal{E}(\ell) = \min_{\varrho} \|\mathbf{R}_v(\ell) - \varrho\hat{\mathbf{R}}_v(\ell)\|_F \tag{50}$$

which is expressed in dB as $20\log_{10}\mathcal{E}(\ell)$, where $\mathbf{R}_v(\ell)$ is defined in (3) and $\varrho$ is used as in [41], to make the metric independent of an arbitrary scaling factor. This independence of scale is appropriate in the context of MVDR beamforming, which is also independent of scale, and allows direct comparison with the fixed-scale model covariance matrices used as baselines, described below. The mean NCM estimation error, $\bar{\mathcal{E}}$, reported in Experiment 3, is given by

$$\bar{\mathcal{E}} = \frac{1}{L}\sum_{\ell\in\mathcal{L}}\mathcal{E}(\ell) \tag{51}$$

where $\mathcal{L}$ is the set of $\ell$ for which $\mathcal{H}_0(\ell) = 0$ and $L$ is the size of $\mathcal{L}$.

The NCM estimation performance is also evaluated in Experiment 4 in terms of the resulting MVDR beamformer performance. The output of the beamformer, $Z(\ell)$, is

$$Z(\ell) = \mathbf{w}^H(\ell)\mathbf{v}(\ell) \tag{52}$$

where the weights, $\mathbf{w}(\ell)$, are obtained as [42]

$$\mathbf{w}(\ell) = \frac{\hat{\mathbf{R}}_{\mathbf{v}}^{-1}(\ell)\mathbf{h}(\Omega)}{\mathbf{h}(\Omega)^H\hat{\mathbf{R}}_{\mathbf{v}}^{-1}(\ell)\mathbf{h}(\Omega)} \tag{53}$$

where $\hat{\mathbf{R}}_{\mathbf{v}}^{-1}(\ell)$ is the inverse of the estimated NCM and $\mathbf{h}(\Omega) = [\,h_1(\Omega)\ \ldots\ h_Q(\Omega)\,]^T$ is the steering vector. The look direction is fixed in array coordinates to $\Omega = (90^\circ, 0^\circ)$, that is, towards the front of the array. The beamformer performance is characterized by the noise reduction, $\gamma$, defined for this purpose as

$$\gamma = \frac{1}{LQ}\sum_{\ell\in\mathcal{L}}\frac{\mathbf{v}^H(\ell)\mathbf{v}(\ell)}{|Z(\ell)|^2}. \tag{54}$$

The oracle beamformer, designed using the ground truth NCM, defines the best case performance. The excess noise level, $\Delta\gamma$, of a beamformer is the amount by which the noise power at the output of a beamformer exceeds that of the oracle beamfomer.

The proposed method is compared to the following three common noise covariance models and estimation approaches.

*(a) Spatially white model (*e.g.*, [4], [5]):* The variance of the noise is assumed to be the same at each microphone, as in (37). Since $\mathcal{E}$ is scale invariant, without loss of generality, the diagonal matrix can be reduced to an identity matrix, $\hat{\mathbf{r}}_{\text{white}} = \overrightarrow{\mathbf{I}}$.

*(b) Spherically isotropic model (*e.g.*, [6], [7]):* For a spherically isotropic noise field, i.e., $\underline{s}(\underline{\Omega})$ is constant over $\underline{\Omega}$, (32) reduces to

$$\hat{\mathbf{r}}_{\text{iso}}(\ell) \propto \overrightarrow{\widetilde{\mathbf{H}}^H\widetilde{\mathbf{H}}}. \tag{55}$$

As for the proposed method, it is assumed that the array manifold is known.

*(c) Recursive smoothing approach (*e.g.*, [18]):* If $\mathcal{H}_0(\ell) = 1$, the estimated NCM is updated on each frame as

$$\begin{aligned}\hat{\mathbf{r}}_{\text{smooth}}(\ell) = {}& (1-\alpha(\ell))\,\hat{\mathbf{r}}_{\text{smooth}}(\ell-1)\\ &+ \alpha(\ell)\overrightarrow{\mathbf{v}(\ell)\mathbf{v}^H(\ell)}\end{aligned} \tag{56}$$

where $\alpha(\ell)$ is the smoothing factor which controls the trade off between tracking changes in $\mathbf{r}_v(\ell)$ and the variance of the estimate. Results are shown for $\alpha \in \{1,5,10,50,100\}\times 10^{-3}$. If $\mathcal{H}_0(\ell) = 0$, the estimated NCM is not updated, i.e.,

$$\hat{\mathbf{r}}_{\text{smooth}}(\ell) = \hat{\mathbf{r}}_{\text{smooth}}(\ell-1). \tag{57}$$

### C. Experiment 1 - Effect of Orientation Constraints

The proposed method estimates the SH coefficients of a function which varies over the surface of a sphere. However, the sensors in the unrotated microphone array all lie in a horizontal plane. The experiment investigates the extent to which the accuracy of the model parameter estimation depends on the spatial diversity of sampling introduced by array rotation.

Three rotation sequences are considered, each consisting of 50 distinct array orientations. For sequence 1, denoted 'yaw only', the yaw is incremented in equal steps from $0^\circ$ to $1080^\circ$, i.e., 3 full rotations, while pitch and roll are fixed at $0^\circ$. Sequence 2, denoted 'constrained', follows the same yaw rotations as sequence 1, but pitch and roll are drawn randomly from normal distributions with standard deviations $\sqrt{20}^\circ$ and $\sqrt{10}^\circ$, respectively. This introduces some vertical diversity into the microphone positions over a range that is believed to be representative for a head-mounted array. In sequence 3, denoted 'unconstrained', yaw, pitch and roll rotations for each of the 50

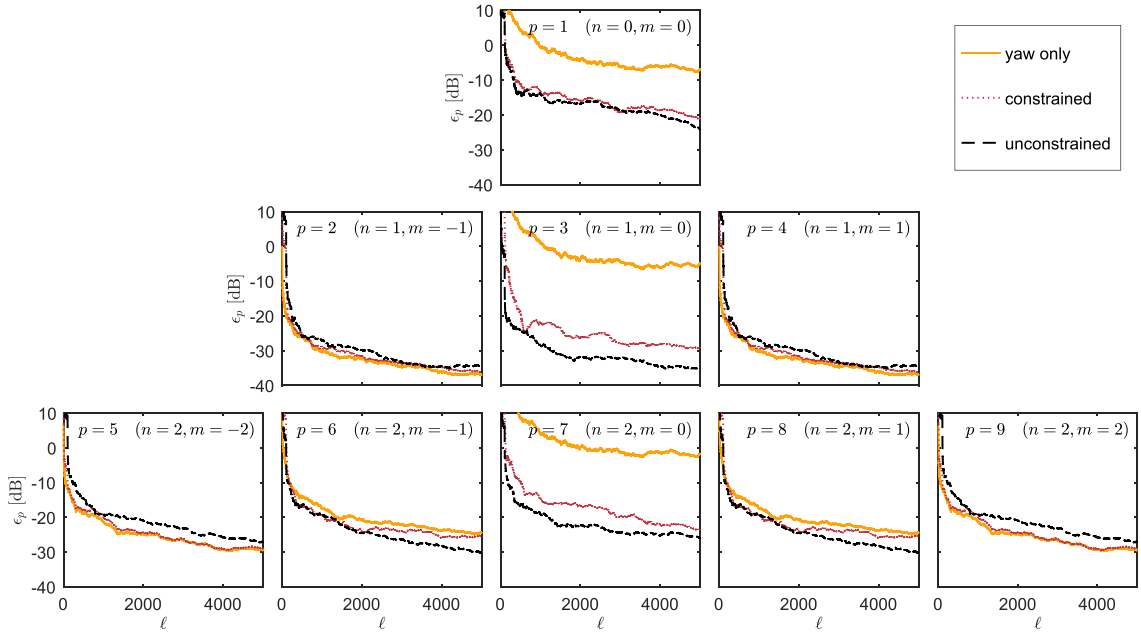



Fig. 4. Convergence of model parameters, averaged over 20 directionally-uncorrelated fields with equally spaced $N_s = 2$ cardioid power distribution, for three different rotation sequences, where $n$ is the SH order and $|m| \leq n$ is the SH degree.

orientations are drawn uniformly from the interval $[0°, 360°)$. This represents unconstrained rotation of the microphone array. There is no desired source activity so $\mathcal{H}_0(\ell) = 1$ throughout. For each orientation, the proposed method is allowed to adapt for 100 frames. Results are averaged over the 20 different noise fields.

Fig. 4 shows the error in the estimated model parameters, $\epsilon_p(\ell)$. The convergence is comparable across rotation sequences for all SH coefficients except those of degree 0, where 'unconstrained' rotation converges to the lowest error, $-24.0$, $-35.1$ and $-25.9$ dB for $n = 0, 1$ and 2, respectively, compared to $-21.0$, $-29.1$ and $-23.6$ dB for 'constrained' and $-7.0$, $-4.9$ and $-1.9$ dB for 'yaw only' rotation. These results show that introducing vertical diversity in the microphone positions through array rotations helps to disambiguate between the SH functions of degree 0, which are symmetrical around the $z$-axis.

Fig. 5 shows the Frobenius norm of the scale-invariant NCM estimation error, $\mathcal{E}(\ell)$, for the three sequences. Despite the differences in the parameter estimates seen in Fig. 4, the error in the resulting covariance matrix estimates are within 1dB of each other. This is a consequence of the underdetermined nature of the estimation problem. That converging on the correct solution for the model parameters is not a necessary condition for achieving a good estimate of the NCM suggests that the method can yield useful results even when the array rotation is constrained.

### D. Experiment 2 - Effect of Rotation During Source Activity

In the remaining experiments the rotation sequence and speech absence state, $\mathcal{H}_0(\ell)$, are chosen to reflect a situation in which array rotation is in response to desired source activity. Whenever $\mathcal{H}_0(\ell) = 1$ the proposed method and the conventional method use noise-only microphone signals to update the estimated NCM. Whenever $\mathcal{H}_0(\ell) = 0$, the conventional method does not update but, in contrast, the proposed method uses the previously estimated NPD to estimate the NCM from the current array orientation.

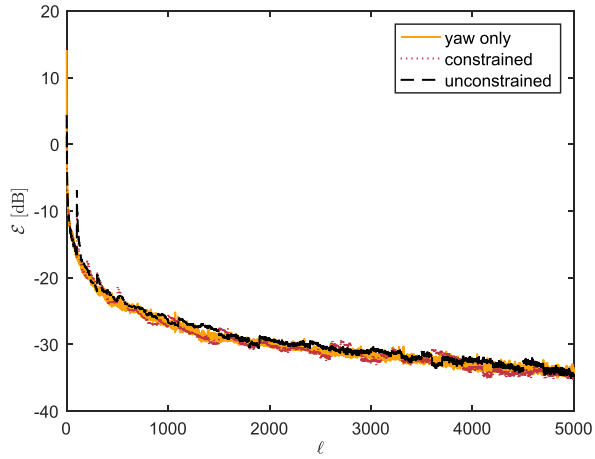



Fig. 5. Convergence of estimated NCM, averaged over 20 directionally-uncorrelated fields with equally spaced $N_s = 2$ cardioid power distribution, for three different rotation sequences.

The first four orientations have deterministic yaw angles, $\{0°, 30°, 60°, 90°\}$ while pitch and roll are stochastic, as in sequence 2. In the final orientation, roll, pitch and yaw are all $0°$. Each orientation is held for 250 frames. For $\ell \leq 950$, $\mathcal{H}_0(\ell) = 1$ while for $\ell > 950$, $\mathcal{H}_0(\ell) = 0$.

Fig. 6 shows the convergence of $\mathcal{E}(\ell)$ for the conventional RS method, denoted 'RS: $\alpha$', and the proposed method, denoted 'EWLS: $(1 - \lambda)$', over a range of time constants. As a reference, $\mathcal{E}(\ell)$ is also shown for $\hat{\mathbf{r}}_{\mathrm{white}}$, denoted 'White' and $\hat{\mathbf{r}}_{\mathrm{iso}}$, denoted 'Sph iso'.

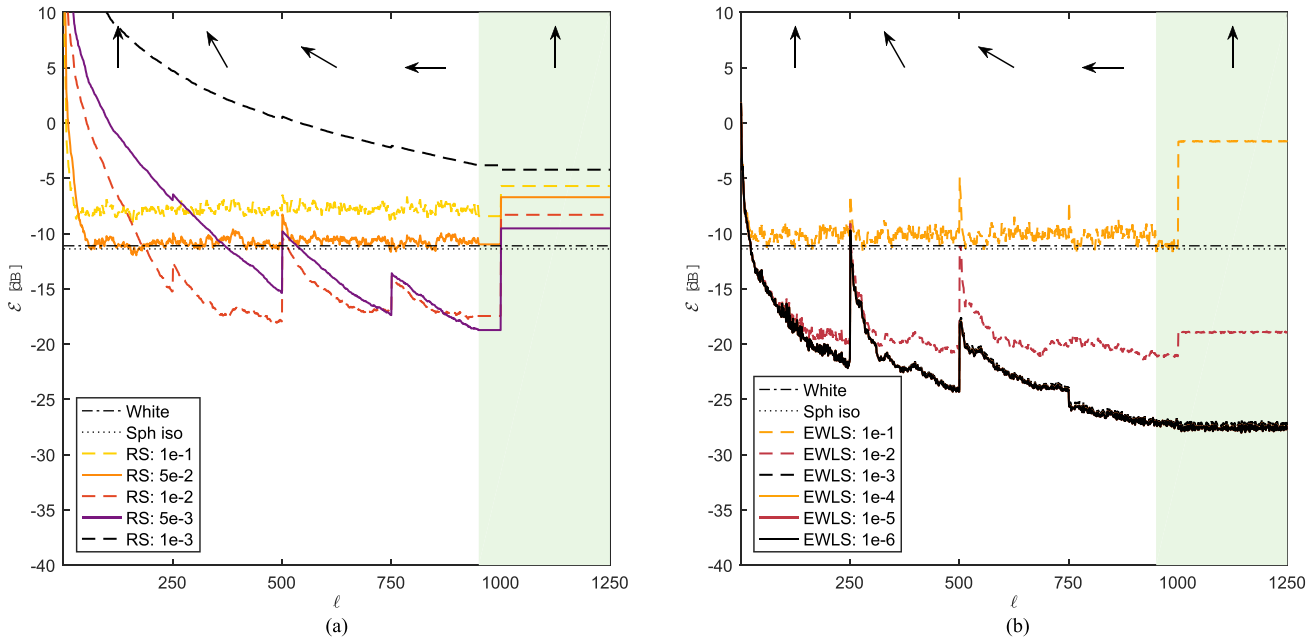



Fig. 6. Noise covariance error over time for (a) recursive smoothing (recursive smoothing (RS)) and (b) proposed method (EWLS) with different time constants. Error using signal independent models, 'White' and 'Sph iso' also shown for comparison. Shaded region indicates $\mathcal{H}_0(\ell) = 0$. Overlayed arrows indicate yaw component of array rotation, which is stepped every 250 frames in sequence $\{0°, 30°, 60°, 90°, 0°\}$.

For RS, the minimum error is obtained with $\alpha = 5 \times 10^{-3}$. In this case the lowest value of $\mathcal{E}(\ell)$ acheived is $-18.7$ dB and on each orientation change there is a large increase in $\mathcal{E}(\ell)$ before quickly converging again. The problem with the conventional method is clearly demonstrated at $\ell = 1001$, where the orientation changes, but because $\mathcal{H}_0(\ell) = 0$ the estimate cannot be updated. The error in the estimated noise covariance during desired source activity is $-9.5$ dB which is larger than for the straightforward model-based estimates which achieve $-11.1$ dB and $-11.4$ dB for spatially white and spherically isotropic models, respectively.

Varying the value of $\alpha$ degrades performance for both states of $\mathcal{H}_0(\ell)$; a smaller value means that convergence is too slow whilst a larger value means that instantaneous variations in the noise level are tracked, leading to an overall higher error.

For the proposed method with $(1 - \lambda) \leq 1e - 3$, convergence is insensitive to the choice of $\lambda$ with no visible difference over three orders of magnitude. The minimum value of $\mathcal{E}(\ell)$ is $-27.1$ dB at $\ell = 950$, which is consistent with the value observed in Experiment 1 at $\ell = 950$, despite the rotation sequence containing fewer distinct orientations. Crucially, at $\ell = 1001$ when $\mathcal{H}_0(\ell) = 0$ and the orientation changes there is no increase in $\mathcal{E}(\ell)$. The proposed method therefore achieves 18 dB lower error than the RS approach.

The fundamental difference between the two approaches is that the proposed method adapts to the properties of the noise field in world coordinates and so the choice of $\lambda$ depends only on how quickly the NPD changes whereas the RS method must adapt to changes in the observation of the noise field through the microphone signals in array coordinates and so $\alpha$ must be chosen to allow for more rapid adaptation. It should be noted that the piecewise-constant trajectories used in this evaluation intentionally emphasize this dependence on $\alpha$. Smoothly varying trajectories would not exhibit the same spikes in estimation error during speech absence. Nevertheless, the ability to track changes in the NCM depends on the speed of rotation rather than the smoothness.

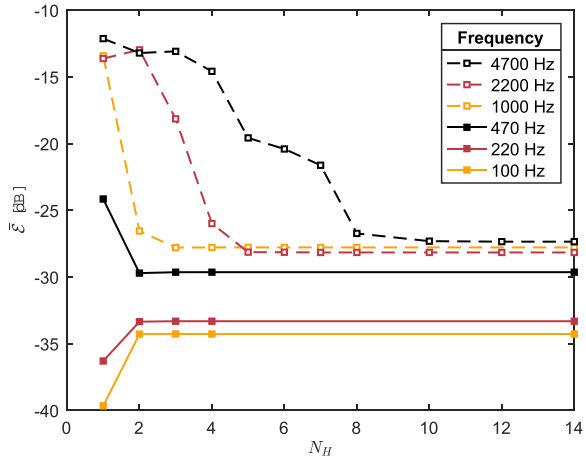



Fig. 7. Effect of truncation order of PWD, $N_h$, on noise covariance estimation at different frequencies.

### *E. Experiment 3 - Effect of Array Manifold Sampling Density*

The proposed method requires the array manifold to be sampled in azimuth and inclination so that its SFT can be computed as in (7). The maximum SH order used to represent the array manifold, $N_h$, determines the number of directions which must be sampled which, depending on the sampling scheme used, is lower bounded at $I \geq (N_h + 1)^2$. Fig. 7 shows $\bar{\mathcal{E}}$ from (51) as a

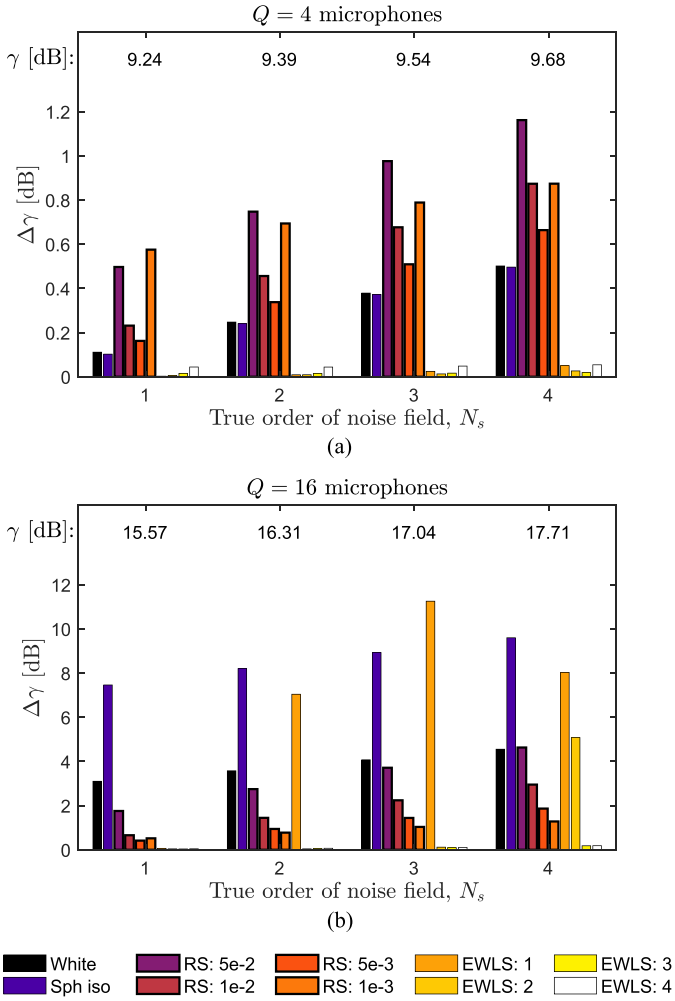



Fig. 8. Excess noise power, $\Delta\gamma$, of MVDR beamformer with (a) $Q = 4$ and (b) $Q = 16$ with different noise covariance estimation methods, averaged over 20 $N_s$-order cardioid-shaped noise fields. Results are grouped by the true NPD order, $N_s$. The noise reduction, $\gamma$, obtained from an oracle MVDR beamformer is shown above each group. Labels for White, Sph iso and RS methods are as for Fig. 6. Labels for the proposed method with different estimation orders are denoted 'EWLS: $N_{\hat{s}}$'.

function of $N_h$ at different frequencies and gives an indication of the number of directions which must be measured. For frequencies $\leq 1$ kHz, $\bar{\mathcal{E}}$ with $N_h = 4$ is as good as with $N_h = 14$, suggesting that as few as 25 measurements are sufficient. At higher frequencies, the sampling requirements increase rapidly with $N_h = 10$ required for best-case performance at 4700 Hz.

### *F. Experiment 4 - Effect of Model Mismatch and Number of Microphones on Beamforming*

One possible application of the proposed NCM estimation method is in MVDR beamforming. In this experiment, the noise reduction performance of MVDR beamformers based on estimated NCMs is presented. In particular, the effect of model mismatch between the estimation order, $N_{\hat{s}}$, and the true NPD order, $N_s$, is investigated. Fig. 8(a) shows the excess noise, $\Delta\gamma$, for the same $Q = 4$ microphone array as in Experiments 1–3. As the noise field becomes more directional there is a small increase in $\gamma$ obtained using the oracle beamformer, 9.2 dB to 9.7 dB. The proposed method achieves the best performance of $\Delta\gamma \leq 0.05$ dB over all $N_s$ and $N_{\hat{s}}$. This implies that choosing $N_{\hat{s}} \neq N_s$ does not have a detrimental effect on the beamforming performance.

In contrast, the RS method with best-case smoothing parameter, $\alpha = 5 \times 10^{-3}$, achieves $\Delta\gamma = 0.16$ dB for $N_s = 1$, rising to $\Delta\gamma = 0.66$ dB for $N_s = 4$. That is, as the sound field becomes more directional the 0.5 dB improvement in the oracle beamformer's noise reduction is not matched when using the RS estimate of the NCM. Using either a spatially white or spherically isotropic model of the noise field acheives 0.05 dB to 0.16 dB more noise reduction than the best RS estimate.

Fig. 8(b) shows the beamforming performance when the number of microphones is increased to 16. The oracle beamformer can better match the beam pattern to the power distribution of the noise field than with 4 microphones and so more noise reduction is acheived (15.57 dB to 17.71 dB), with the benefit increasing with $N_s$. With the proposed method, provided $N_{\hat{s}} \geq N_s$, performance very close to that of the oracle beamformer is acheived, with $\Delta\gamma \leq 0.18$ dB. With RS the best-case excess noise rises from 0.43 dB for $N_s = 1$ to 1.28 dB for $N_s = 4$. Since with $Q = 16$ there are more degrees of freedom, the effect of undermodelling is more severe than for $Q = 4$. This is true for the spatially white model – which does not account for the cross-terms in the NCM or the intersensor variations in noise power – the spherically isotropic model – which does not account for the directional variation in the NPD and ignores the sensor noise – and for the proposed method with $N_{\hat{s}} < N_s$. The spatially white assumption is the most robust since, by not modelling the interchannel correlations, it also does not attempt to exploit them in the noise reduction. This leads to 3.10 dB to 4.55 dB more residual noise than the oracle beamformer. In contrast, the effect of errors due to the spherically isotropic assumption and the undermodelled proposed method, when inverted in (53), degrade the beamformer by 7.46 dB to 9.60 dB and 7.04 dB to 11.25 dB, respectively.

## VII. Conclusion

A model for non-isotropic directionally-uncorrelated noise has been proposed based on a SH decomposition of the sound field. An analytical expression for the noise covariance matrix is obtained directly from the proposed model using knowledge of the array manifold and the array orientation. An algorithm for estimating the parameters of the proposed model has been proposed and validated on simulated noise fields with realistic levels of microphone sensor noise. The approach is particularly suited to situations in which changes in array orientation in response to and during desired source activity are expected. In this context, the proposed method achieves 18 dB lower error in the estimated noise covariance matrix than the conventional recursive averaging approach and noise reduction which is within 0.05 dB of an oracle beamformer using the ground truth noise covariance matrix.

## APPENDIX

In (21) and (23), the acoustic NCM is related to the covariance between the SH coefficients of the sound field. We here derive the relationship between these SH covariance terms and the proposed model in (25).

The element of $\mathbf{R}_{\mathbf{a}}^{H}$ at row $p''$ and column $p'$ is given by $\mathbb{E}\left\{a_{p'}(\ell)a_{p''}^{*}(\ell)\right\}^{*} = \mathbb{E}\left\{a_{p'}^{*}(\ell)a_{p''}(\ell)\right\}$. Substituting the SFT (5) of $a(\ell,\Omega)$, i.e.,

$$a_p(\ell) = \int_{\Omega\in S^2} a(\ell,\Omega)Y_p^*(\Omega)d\Omega \tag{58}$$

gives

$$\mathbb{E}\left\{a_{p'}^{*}(\ell)a_{p''}(\ell)\right\} \tag{59}$$

$$= \mathbb{E}\left\{\left[\int_{\Omega\in S^2} a(\ell,\Omega)Y_{p'}^{*}(\Omega)d\Omega\right]^{*} \times \int_{\Omega'\in S^2} a(\ell,\Omega')Y_{p''}^{*}(\Omega')d\Omega'\right\} \tag{60}$$

$$= \mathbb{E}\left\{\int_{\Omega\in S^2} a^{*}(\ell,\Omega)Y_{p'}(\Omega)d\Omega \times \int_{\Omega'\in S^2} a(\ell,\Omega')Y_{p''}^{*}(\Omega')d\Omega'\right\} \tag{61}$$

$$= \int_{\Omega\in S^2}\int_{\Omega'\in S^2} \mathbb{E}\left\{a(\ell,\Omega')a^{*}(\ell,\Omega)\right\}Y_{p'}(\Omega)Y_{p''}^{*}(\Omega')d\Omega d\Omega'. \tag{62}$$

Expressing (24) in array coordinates gives

$$\mathbb{E}\left\{a(\ell,\Omega)a^{*}(\ell,\Omega')\right\} = s(\Omega)\delta(\Omega-\Omega') \tag{63}$$

which substituted into (62) gives

$$\mathbb{E}\left\{a_{p'}^{*}(\ell)a_{p''}(\ell)\right\} \tag{64}$$

$$= \int_{\Omega\in S^2} s(\Omega)Y_{p'}(\Omega)Y_{p''}^{*}(\Omega)d\Omega \tag{65}$$

$$= \int_{\Omega\in S^2} Y_{p'}(\Omega)Y_{p''}^{*}(\Omega)\mathbf{y}^{T}(\Omega)\mathbf{D}(\Lambda^{-1})\underline{\mathbf{s}}d\Omega \tag{66}$$

where the last line uses (27). Using the identity

$$G_{p,p',p''} = \int_{\Omega\in S^2} Y_p(\Omega)Y_{p'}(\Omega)Y_{p''}^{*}(\Omega)d\Omega \tag{67}$$

and $\mathbf{g}_{p',p''} = \left[\, G_{1,p',p''} \;\ldots\; G_{P_s,p',p''} \,\right]^{T}$, (66) is written

$$\mathbb{E}\left\{a_{p'}^{*}(\ell)a_{p''}(\ell)\right\} = \mathbf{g}_{p',p''}^{T}\mathbf{D}(\Lambda^{-1})\underline{\mathbf{s}}. \tag{68}$$

The conjugate of (68) gives the element of $\mathbf{R}_{\mathbf{a}}$ at row $p'$ and column $p''$. Since $G_{p,p',p''}$ is real-valued, and $\mathbf{D}^{*}(\Lambda^{-1}) = \mathbf{D}^{T}(\Lambda)$, this can be expressed as

$$\mathbb{E}\left\{a_{p'}(\ell)a_{p''}^{*}(\ell)\right\} = \mathbf{g}_{p',p''}^{T}\mathbf{D}^{T}(\Lambda)\underline{\mathbf{s}}^{*}. \tag{69}$$

## REFERENCES


[1] J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[2] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[3] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 39–60.

[4] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New York, NY, USA, Apr. 1988, pp. 2578–2581.

[5] R. Zelinski, "Noise reduction based on microphone array with LMS adaptive post-filtering," *IEE Electron. Lett.*, vol. 26, no. 24, pp. 2036–2581, Nov. 1990.

[6] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[7] S. Leukimmiatis and P. Maragos, "Optimum post-filter estimation for noise reduction in multichannel speech processing," in *Proc. Eur. Signal Process. Conf.*, Sep. 2006, pp. 1–5.

[8] S. Braun, D. P. Jarrett, J. Fischer, and E. A. P. Habets, "An informed spatial filter for dereverberation in the spherical harmonic domain," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 669–673.

[9] M. Jeub, C. M. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc. Eur. Signal Process. Conf.*, Barcelona, Spain, 2011, pp. 1347–1351.

[10] O. Thiergart, G. D. Galdo, and E. A. P. Habets, "Diffuseness estimation with high temporal resolution via spatial coherence between virtual first-order microphones," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2011, pp. 217–220.

[11] O. Thiergart, G. D. Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J. Acoust. Soc. Amer.*, vol. 132, no. 4, pp. 2337–2346, 2012.

[12] O. Thiergart, G. D. Galdo, and E. A. P. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2012, pp. 309–312.

[13] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *Proc. IEEE Conv.. Elect. Electron. Israel*, Eilat, Israel, Nov. 2012, pp. 1–5.

[14] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 1006–1018, Jun. 2015.

[15] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids," in *Proc. Eur. Signal Process. Conf.*, Sep. 2014, pp. 61–65.

[16] A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen, and J. Jensen, "Multi-channel PSD estimators for speech dereverberation—A theoretical and experimental comparison," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 91–95.

[17] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[18] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 295–299.

[19] U. Hadar, T. J. Steiner, and F. C. Rose, "Head movement during listening turns in conversation," *J. Nonverbal Behav.*, vol. 9, no. 4, pp. 214–228, Dec. 1985.

[20] M. Zohourian, A. Archer-Boyd, and R. Martin, "Multi-channel speaker localization and separation using a model-based GSC and an inertial measurement unit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5615–5619.

[21] M. Zohourian and R. Martin, "Binaural speaker localization and separation based on a joint ITD/ILD model and head movement tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 430–434.

[22] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. London, U.K.: Academic Press, 1999.

[23] V. Tourbabin, H. Barfuss, B. Rafaely, and W. Kellermann, "Enhanced robot audition by dynamic acoustic sensing in moving humanoids," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, Apr. 2015, pp. 5625–5629.

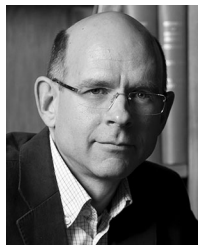
[24] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *J. Acoust. Soc. Amer.*, vol. 127, pp. 823–828, 2010.

[25] J. Sheaffer, S. Villeval, and B. Rafaely, "Rendering binaural room impulse responses from spherical microphone array recordings using timbre correction," in *Proc. EAA Joint Symp. Auralization Ambisonics*, Apr. 2014, pp. 81–85.

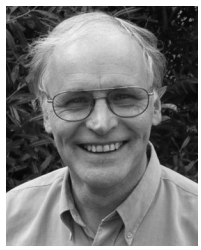
[26] J. Sheaffer and B. Rafaely, "Equalization strategies for binaural room impulse response rendering using spherical arrays," in *Proc. IEEE Conv. Elect. Electron. Eng. Israel*. Eilat, Israel, 2014, pp. 1–5.
[27] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Microphone array measurement system for analysis of directional and spatial variations of sound fields," *J. Acoust. Soc. Amer.*, vol. 112, no. 5, pp. 1980–1991, 2002.
[28] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *J. Acoust. Soc. Amer.*, vol. 116, no. 4, pp. 2138–2148, Oct. 2004.
[29] J. M. Rigelsford and A. Tennant, "Acoustic imaging using a volumetric array," *Appl. Acoust.*, vol. 67, no. 7, pp. 680–688, Jul. 2006.
[30] B. Rafaely, "Spherical microphone array with multiple nulls for analysis of directional room impulse responses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 281–284.
[31] N. Epain and C. T. Jin, "Spherical Harmonic Signal Covariance and Sound Field Diffuseness," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1796–1807, Oct. 2016.
[32] V. Tourbabin and B. Rafaely, "Analysis of distortion in audio signals introduced by microphone motion," in *Proc. 24th Eur. Signal Process. Conf.*, Aug. 2016, pp. 998–1002.
[33] B. Rafaely, *Fundamentals of Spherical Array Processing* (Springer Topics in Signal Processing). Berlin, Germany: Springer-Verlag, 2015.
[34] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing* (Springer Topics in Signal Processing). Berlin, Germany: Springer, 2017.
[35] M. Brookes, "The Matrix Reference Manual," Imperial College London, Website, 1998.[Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/relation.html#Kronecker
[36] M. E. Rose, *Elementary Theory of Angular Momentum*. New York, NY, USA: Wiley, Dec. 1957.
[37] H. Sun, E. Mabande, K. Kowalczyk, and W. Kellermann, "Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing," *J. Acoust. Soc. Amer.*, vol. 131, pp. 2828–2840, 2012.
[38] A. H. Moore, C. Evers, and P. A. Naylor, "Direction of arrival estimation in the spherical harmonic domain using subspace psuedointensity vectors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2016. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2016.2613280
[39] S. Haykin, *Adaptive Filter Theory*, 4th ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.
[40] C. T. Jin, N. Epain, and A. Parthy, "Design, optimization and evaluation of a dual-radius spherical microphone array," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 193–204, Jan. 2014.
[41] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. Berlin, Germany: Springer-Verlag, 2010.
[42] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

**Alastair H. Moore** (M'13) received the M.Eng. degree in Electronic Engineering with Music Technology Systems in 2005 and the Ph.D. degree in 2010, both from the University of York, York, U.K. He is currently a Postdoctoral Researcher with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. He spent 3 years as a Hardware Design Engineer for Imagination Technologies PLC designing digital radio and networked audio consumer electronics products. In 2012, he joined Imperial College, where he has contributed to a series of projects in the field of speech and audio processing applied to voice over IP, robot audition, and hearing aids. Particular topics of interest include microphone array signal processing, modeling and characterization of room acoustics, dereverberation, and spatial audio perception.

**Wei Xue** (M'16) received the B.Eng. degree in Automatic Control from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He has been a Research Scientist with JD AI Research Laboratory, Beijing, China, since November 2018. From August 2015 to September 2018, he was first a Marie Curie Experienced Researcher and then a Research Associate with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K. He was a visiting scholar with the Universit de Toulon, in July 2015, and KU Leuven, in September 2016. His research interest focuses on microphone arrays based speech signal processing, including speech enhancement, sound source localization, and blind system identification.



**Patrick A. Naylor** (M'89–SM'07) received the B.Eng. degree in Electronic and Electrical Engineering from the University of Sheffield, Sheffield, U.K., and the PhD. degree from Imperial College London, London, U.K. He is currently a member of academic staff in the Department of Electrical and Electronic Engineering, Imperial College London. His research interests include speech, audio, and acoustic signal processing. He has worked in particular on speech dereverberation, including blind multichannel system identification and equalization, as well as acoustic echo control, nonintrusive speech quality estimation, single and multichannel speech enhancement, and speech production modeling with a focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several collaborative links with industry. He is the President of the European Association for Signal Processing (EURASIP) and was formerly the Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing. He has served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and is currently a Senior Area Editor for the IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING.



**Mike Brookes** (M'88) received the B.A. degree in Mathematics from Cambridge University, Cambridge, U.K., in 1972. He is currently a Reader (Associate Professor) in Signal Processing with the Department of Electrical and Electronic Engineering, Imperial College London. He worked with the Massachusetts Institute of Technology and, briefly, the University of Hawaii before returning to the U.K. and joining Imperial College in 1977. Within the area of speech processing, he has concentrated on the modeling and analysis of speech signals, the extraction of features for speech and speaker recognition, and on the enhancement of poor-quality speech signals. He is the primary author of the VOICEBOX speech processing toolbox for MATLAB. Between 2007 and 2012, he was the Director of the Home Office sponsored Centre for Law Enforcement Audio Research (CLEAR), which investigated techniques for processing heavily corrupted speech signals. He is currently a Principal Investigator of the E-LOBES project that seeks to develop environment-aware enhancement algorithms for binaural hearing aids.